Abstract

An evaluator system accepts input textual messages in unknown languages and assesses which character sets, corresponding to languages, matches that message. Textual messages whose individual characters are encoded in 16 bit Unicode or other universal format are parsed, and character sets which can express each character and the accumulated correspondence is logged. When the character sets against which the message is being tested only provide partial matches, the invention can determine which offers the best fit, including by means of a weighting function. The evaluation technology of the invention can be applied to multipart documents, and to search engines and indices. Documents can be indexed according to assigned character sets, and quary strings matched to indices according to language.